Recent advances in inferring viral diversity from high-throughput sequencing data

Virus Research 239 (2017) 17–32

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Review

Recent advances in inferring viral diversity from high-throughput sequencing data

Susana Posada-Cespedes[a,b], David Seifert[a,b], Niko Beerenwinkel[a,b,*]

[a] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[b] SIB, Basel, Switzerland

ARTICLE INFO

Article history:
Received 24 June 2016
Received in revised form
23 September 2016
Accepted 24 September 2016
Available online 28 September 2016

Keywords:
Viral quasispecies
Genetic diversity
Haplotype reconstruction
Next-generation sequencing

ABSTRACT

Rapidly evolving RNA viruses prevail within a host as a collection of closely related variants, referred to as viral quasispecies. Advances in high-throughput sequencing (HTS) technologies have facilitated the assessment of the genetic diversity of such virus populations at an unprecedented level of detail. However, analysis of HTS data from virus populations is challenging due to short, error-prone reads. In order to account for uncertainties originating from these limitations, several computational and statistical methods have been developed for studying the genetic heterogeneity of virus population. Here, we review methods for the analysis of HTS reads, including approaches to local diversity estimation and global haplotype reconstruction. Challenges posed by aligning reads, as well as the impact of reference biases on diversity estimates are also discussed. In addition, we address some of the experimental approaches designed to improve the biological signal-to-noise ratio. In the future, computational methods for the analysis of heterogeneous virus populations are likely to continue being complemented by technological developments.

Contents

* Corresponding author
E-mail address: niko.beerenwinkel@bsse.ethz.ch (N. Beerenwinkel).

http://dx.doi.org/10.1016/j.virusres.2016.09.016

1. Introduction

The evolutionary dynamics of RNA viruses, such as the human immunodeficiency virus (HIV), the hepatitis C virus (HCV), or influenza virus, is characterized by high mutation rates, short generation times and large population sizes (Duffy et al., 2008). Under these conditions, a collection of non-identical but related genetic variants is able to co-exist within the host. This ensemble of variants has been referred to as a viral quasispecies (Domingo et al., 2005; Lauring and Andino, 2010). The term quasispecies was first used by Eigen and Schuster (1977), in the context of their work on molecular evolution (Eigen and Schuster, 1978, 1978). The quasispecies model was introduced by means of a theoretical framework using chemical kinetics to describe the mutation and selection processes governing the evolution of self-replicating macromolecules. In virology, the quasispecies model has been adopted to describe the evolutionary dynamics of RNA viruses at the population level (Nowak, 1992; Domingo and Holland, 1997).

Mutation and selection are one of the driving forces of evolution in RNA viruses. Largely due to the lack of proof-reading capability of the RNA polymerases (i.e., RNA-dependent RNA polymerase and RNA-dependent DNA polymerase or reverse transcriptase), RNA viruses exhibit high mutation rates (Duffy et al., 2008). For instance, the mutation rate of HIV-1 is on the order of 10^{-5} substitutions per position per generation (Duffy et al., 2008; Mansky and Temin, 1995). As a consequence of these high mutation rates, new viral strains are produced in every replication cycle by means of point mutations, insertions and deletions. Another common source of variability in RNA viruses is recombination. A recombination event can take place when at least two different viral strains infect the same cell, giving rise to a new strain which is a mosaic of its progenitors. On the other hand, selective pressures act upon the virus population as a whole, shaping the distribution of viral strains. For instance, in response to changing environments, the virus population quickly adapts by selecting preexisting strains with higher fitness (Bonhoeffer and Nowak, 1997). As a result, one or few viral strains dominate, surrounded by a large cloud of low-frequency variants.

The heterogeneous mixture of viral strains appears to confer numerous advantages to the virus population, including the ability to escape from the host's immune response (Nowak et al., 1991; Kuroda et al., 2010; Woo and Reifman, 2012; Borucki et al., 2013), and the development of resistance to vaccines (Gaschen et al., 2002) and antiviral drugs (Johnson et al., 2008). Furthermore, the existence of different viral strains has significant implications for viral pathogenesis, virulence, persistence and disease progression, and likely contributes to tissue tropism (Vignuzzi et al., 2006; Tsibris et al., 2009; Rozera et al., 2014). The robust adaptability featured by RNA viruses, which is related to their genetic heterogeneity is, thus, of clinical relevance. In fact, many of the infectious diseases which have jeopardized and still are a threat to public health are caused by RNA viruses, including HIV, HCV, Influenza virus, Ebola virus and Zika virus.

Before the establishment of HTS technologies, Sanger sequencing was the method of choice for analyzing virus samples. Even today, it remains the gold standard for many clinical applications. However, bulk sequencing only allows for determining the consensus sequence of the virus population. The consensus sequence is an aggregate of all variants within the population. Consequently, it is dominated by highly abundant strains and cannot be used to assess the linkage of mutations in individual variants (Wirden et al., 2005; Zagordi et al., 2010). Further experimental improvements, including isolation of individual viral strains through cloning (Domingo, 2015) or limiting dilutions (Palmer et al., 2005), allow to acquire a better, yet small, sample of the variants within the virus population. This is because these protocols are labor- and time-intensive and, thus, scalability remains a limiting factor.

The sensitivity and scalability issues are progressively being overcome by a set of newer technologies, which allow to produce massive volumes of genomic data in a relatively short time by parallelization of the sequencing reactions. These technologies are collectively referred to as high-throughput sequencing (HTS), massively parallel sequencing (MPS), next-generation sequencing (NGS) or ultra-deep sequencing (UDS). HTS technologies allow an in-depth characterization of the genetic diversity in heterogeneous virus populations by directly sequencing many of the viral strains. Furthermore, provided that the sequencing coverage is sufficiently high, it is possible to detect mutations present in less abundant strains, whereas consensus Sanger sequencing has a 20% detection threshold. However, low-frequency mutations are particularly relevant in the context of drug resistance, since they may facilitate viral adaptation leading to treatment failure (Metzner et al., 2009; Gianella and Richman, 2010; Avidor et al., 2013; Vandenhende et al., 2014). Therefore, studying the genetic diversity of the virus population as a whole is more informative than focusing solely on the dominant viral strains.

HTS technologies have the potential to provide a representative sample of the virus population. However, many HTS platforms generate large amounts of sequencing reads with short read lengths and relatively high error rates. These factors, in conjunction with errors associated with sample preparation (e.g., RNA extraction, reverse transcription and PCR amplification biases), pose computational and statistical challenges for inferring intra-host genetic diversity from HTS reads (Beerenwinkel et al., 2012; McElroy et al., 2014). For instance, many single-nucleotide variants (SNVs) are present at low frequencies and are therefore difficult to distinguish from technical errors. In addition, reconstructing the population structure from sequencing reads is challenging because the number of underlying viral strains is unknown, some of them exist at low relative abundances, and the diversity among strains can be low (i.e., some variants within the population exhibit a small genetic distance). From the technical perspective, reconstruction of full-length haplotypes is challenging because sequencing reads are typically shorter than the viral genome and do not cover the genome or the genetic region of interest uniformly. To this end, recent advances in single-molecule sequencing seem promising, as platforms commercialized by Pacific Biosciences and Oxford Nanopore offer very long reads (>10 kb). However, higher error-rates and lower throughput compared to predecessor HTS platforms still limit applicability of single-molecule sequencers.

Nevertheless, HTS technologies have already proven useful in different fields related to virology, including virus discovery (Cheval et al., 2011), characterization of virus biodiversity found in different environments (also known as virome profiling) (Hurwitz and Sullivan, 2013), estimation of fitness landscapes of viral populations (Seifert et al., 2015), characterization of intra-host virus diversity and population dynamics (Kuroda et al., 2010).

This review is structured as follows. First, we address experimental protocols which have been recently designed to overcome limitations associated with short and error-prone reads (Section 2). These sequencing protocols and accompanying data analysis pipelines have enabled correction of technical errors, as well as reconstruction of viral haplotypes. Next, acknowledging that alignment of sequencing reads is in most cases a prerequisite for subsequent analyses, strategies for read alignment are briefly discussed in Section 3, as well as remaining challenges. Lastly, we describe computational methods developed for studying the genetic diversity of virus populations from HTS reads (Section 4).

2. Experimental protocols for improved error correction and viral diversity estimation

A basic workflow for viral sequencing projects includes sample preparation, choice of sequencing platform, quality assurance, read alignment and identification of genetic variants. The sensitivity of computational methods for variant detection can be improved by identifying and correcting errors introduced during upstream library preparation and sequencing steps. Although, several error-correction algorithms have been designed to improve data quality (Zagordi et al., 2010; Skums et al., 2012), this issue has been also addressed from an experimental design perspective.

One of the first and to date most popular method to remove the overwhelming majority of errors introduced by the PCR step involves the use of short random *k*-mers for tagging sequences. These *k*-mers – in virology more commonly known as primerIDs – are produced as part of the oligonucleotide production step. During the reverse transcription, these specialized primers are used instead of standard RT primers. All produced off-spring molecules will have the same unique tag, which can be employed after sequencing to collapse all reads with the same tag into one consensus sequence (Kinde et al., 2011; Jabara et al., 2011). In this way, most errors are removed via majority voting. The primerID protocol can also be used to estimate the error rate of the PCR branching process, by making a first-order approximation for the number of errors introduced in early cycles (Seifert et al., 2016). If the number of collisions is controlled for, then the primerID protocol possesses the ability to detect failures in the preparative steps, which is crucial for asserting the correctness in clinical diagnostics.

The primerID protocol can however not remove errors introduced in the ligation step of the PCR or during reverse transcription (Seifert et al., 2016). The latter is because templates are only redundantly resampled in the PCR step. In addition, it is known, that laboratory reverse transcriptase (RT) enzymes have higher error rates than common PCR enzymes employed (Seifert et al., 2016). Thus, in turn, most of the errors stem from RT substitutions. The novel circle sequencing (CirSeq) protocol can correct errors in the early phase of the protocol by redundantly incorporating the template onto the DNA template multiple times. This feat is achieved by circularizing the RNA and reverse transcribing it multiple times. PCR mutations can be removed by majority vote, whereas RT mutations can be removed by majority between tandem copies on the same template (Lou et al., 2013). The CirSeq protocol makes the fidelity trade-off by drastically decreasing the realistic fragment size for increased sensitivity. Lastly, both the standard primerID protocol and CirSeq allow for studying viruses only on an amplicon level. While amplicon-based sequencing is relevant for drug resistance loci, it becomes cumbersome and laborious at best to perform whole-genome sequencing in this fashion. An extension of primerIDs to variable-length genomic regions also involves circularizing of the RNA. Instead of transcribing the circularized template multiple times like CirSeq does, the protocol Barcode-directed Assembly for Extra-long Sequences (BAsE-Seq) randomly fragments the circularized DNA, leading to templates with varying lengths (Hong et al., 2014). These variable length templates allow for improved haplotype phasing. Using the BAsE-Seq protocol and data analysis pipeline, it has been possible to reconstruct viral haplotypes of 3 kb in length (Hong et al., 2014).

Finally, while all the protocols provide an attractive path for error correction or phasing of haplotypes beyond local scope, they still do have practical drawbacks. CirSeq and BAsE-Seq both include a circularization, a biochemical step that is kinetically unfavorable and hence inefficient. This in turn will require high input template concentrations, which might be problematic in settings with low viral loads, as in HIV clinical diagnostics (Acevedo and Andino, 2014).

3. Alignment of sequencing reads

A fundamental analysis step in inferring viral diversity from sequence data is read alignment. Sequencing reads can be either mapped to their likely genomic region of origin or assembled *de novo*. The former strategy, dubbed reference-based mapping, is the most widespread choice, although *de novo* assembly of sequencing reads into a consensus sequence has gained increasing interest in recent years (Yang et al., 2013; Mangul et al., 2014; Jayasundara et al., 2015; Malhotra et al., 2016).

3.1. Reference-based mapping

Mapping sequencing reads onto a reference genome relies on the existence of such reference sequence. In fact, reference sequences have been established for many viruses of clinical relevance. However, aligning sequencing reads against a reference sequence may introduce biases (Archer et al., 2010). Assume, e.g., that the virus population contains both, strains which closely resemble the reference sequence, as well as strains which diverge strongly from the reference. The former will be more likely to align successfully against the reference than the latter. Typically, poor-quality alignments are ignored in subsequent analyses. Thus, distantly related sub-populations tend to be underrepresented while estimating viral diversity. A common practice to overcome this issue is to first align the reads to an existing reference genome and then generate a consensus sequence using a position-wise majority vote. Subsequently, reads are aligned to the new consensus sequence (Astrovskaya et al., 2011; Hong et al., 2014). Thereby, it is expected that reads that were not originally mapped, may then be mapped to the consensus sequence. In principle, the process of generating a consensus from mapped reads and realignment can be iteratively repeated until there is no gain in the percentage of mapped reads.

Another challenge in mapping sequencing reads arises from a technical viewpoint. Nowadays, HTS technologies offer sequencing of several million reads in a single experiment. Due to the large volumes of sequencing reads, traditional algorithms for sequence alignment, such as the Needleman–Wunsch and Smith–Waterman algorithms, are computationally very costly. The time complexity for each alignment depends on the product of the length of the reference sequence multiplied by the length of the read. Over the past years, and in order to keep pace with the sequencing throughput, a wide variety of read mappers has been developed. Read mappers rely on different indexing strategies improving upon the quadratic time complexity of traditional algorithms.

Based on indexing techniques implemented by the different read mappers, they can be grouped into two categories (Li and Homer, 2010): algorithms based on (i) hash tables or (ii) prefix/suffix trees. Among software packages belonging to the former group, Stampy (Lunter and Goodson, 2011) and MOSAIK (Lee et al., 2014) have been employed for mapping sequencing reads from mixed samples of virus populations (Mangul et al., 2014; Astrovskaya et al., 2011; Pandit and de Boer, 2014; Cuevas et al., 2015; Zanini et al., 2015). The latter category includes algorithms such as BWA (Li and Durbin, 2009), BWA-SW (Li and Durbin, 2010), Bowtie (Langmead et al., 2009) and Bowtie2 (Langmead and Salzberg, 2012) which employ the Ferragina–Manzini (FM) index (Ferragina and Manzini, 92127) based on the Burrows–Wheeler transform (Burrows and Wheeler, 1994). Several review articles and benchmark studies have been published and may prove useful to guide selection of an adequate tool for a given application (Bao et al., 2011; Fonseca et al., 2012; Caboche et al., 2014).

Run time is a critical aspect, especially when dealing with large eukaryotic genomes, such as the human genome. In theory,

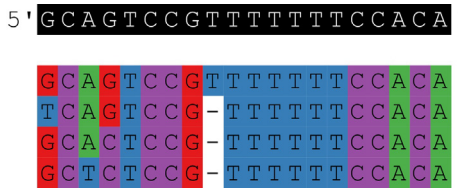

Fig. 1. Placement of gaps in homopolymeric regions. In this hypothetical alignment, four reads are aligned against a window of the reference sequence (black) consisting of 20 nt. The deletion observed in most of the reads is likely to correspond to polymerase slippage errors. However, since the deletion appears to be abundant, it is likely to be picked up as a true variant.

Bowtie (Langmead et al., 2009) and BWA (Li and Durbin, 2009), e.g., align reads in time linear to sequence length, because the data structure of the index only requires querying the read. In practice, and in order to provide efficient solutions for aligning reads, most mappers complement indexing strategies with different heuristics. Nowadays, it is possible to map about 1 gigabases (e.g., 10 million 100 bp reads) per CPU-hour (Langmead et al., 2009; Langmead and Salzberg, 2012). Processing reads from smaller viral genomes can appear as a simpler problem. However, this is only partially true due to a comparatively higher variability of viral genomes. Particularly, heuristics employed to improve run times oftentimes imply a reduction in sensitivity and accuracy. For highly variable viral genomes, inaccurate alignments may result in alignment biases and a non-negligible loss of data, which are propagated into subsequent analysis steps (Archer et al., 2010).

Placement of gaps is yet another challenge in sequence alignment. The most parsimonious alignment, i.e., the alignment with the fewest gaps, is not necessarily the most consistent with the structure of a virus population. There is evidence that supports both frameshift mutations and longer deletions as sources of genetic variation in virus populations (Berthet et al., 1997; Audsley et al., 2010; Guglietta et al., 2010; Reguera et al., 2011; Park et al., 2014). On the other hand, insertions and deletions are not always true sources of variability. The primary source of errors of some sequencing platforms, such as Roche 454, Ion Torrent (Life Technologies) and Pacific Biosciences platforms, are insertions and deletions (collectively referred to as indels) (Loman et al., 2012). However, most read aligners have deficiencies in dealing with indels. Some tools do not support gapped alignments, such as Bowtie (Langmead et al., 2009), and others restrict the number of gaps that are allowed per alignment, such as SOAP (Li et al., 2008) and BWA (Li and Durbin, 2009). More importantly, read aligners supporting gapped alignments tend to place indels in homopolymeric regions either at the beginning or the end of such regions, which leads to calling spurious variants (Fig. 1). To overcome these limitations, a multiple sequence alignment (MSA) approach using statistical models, such as profile hidden Markov models (profile-HMM) (Mount, 2009; Yoon, 2009), could provide a better solution to the read alignment problem. This is because, features shared among related sequences are captured through position-specific scores. If there exist evidence in the population of, e.g., a deletion in a given location, opening a gap in the alignment is allowed with higher probability at this site compared to other positions. Alignment of protein families is an example of a successful application employing profile-HMMs (Eddy, 2003).

3.2. De novo *assembly*

As mentioned earlier, reference biases can be induced by the alignment of the sequencing reads to a reference sequence that highly diverges from the sampled population. In order to circumvent reference biases, a consensus sequence can be assembled *de novo*. Moreover, for virus discovery applications, *de novo* assembly of sequencing reads into a consensus sequences is the only choice.

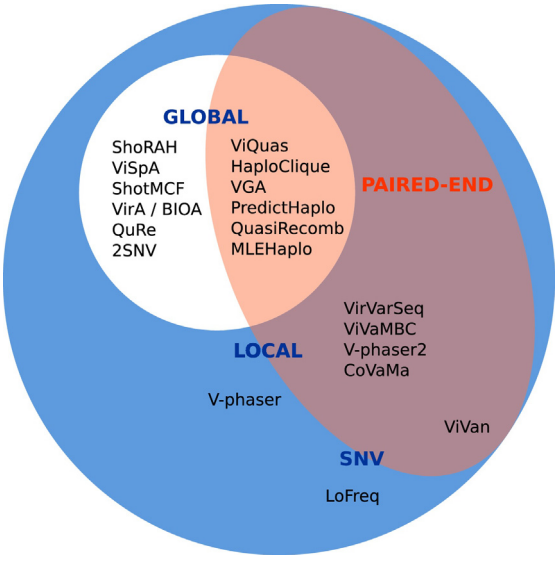

Fig. 2. Available software for assessing viral genetic diversity. Different methods are grouped according to their scope, i.e., SNV calling and local or global reconstruction, and their ability to use information from paired-end reads.

The core concept in *de novo* assembly is to merge overlapping reads into longer stretches of DNA, called contigs, and then merge contigs into scaffolds in order to reconstruct a full-length genome. Genetic heterogeneity of virus populations renders the *de novo* assembly of the reference viral genome more challenging compared to haploid or diploid organisms. However, virus genomes are relatively shorter and generally do not exhibit large repetitive elements compared to, e.g., the human genome.

Several *de novo* assemblers have been tailored to mixed viral samples (Warren et al., 2007; Henn et al., 2012; Yang et al., 2012; Hunt et al., 2015). Among them, the software VICUNA (Yang et al., 2012) constructs consensus sequences of viral genomes by including more heuristics and curating the created reference sequencing using a number of techniques. It creates contigs on the basis of de Bruijn graphs, employing multiple sequence alignments of target genomes to further improve the quality of the contigs. Due to the very small size of the contigs in comparison to large eukaryotic genomes, VICUNA can afford to validate and extend the contigs to possibly full-size genome scales, by filtering likely contaminant reads and improving single base calls using the base pileups from the data.

4. Inference of viral diversity

On the basis of research objectives, genetic diversity of virus populations can be studied at different genomic scales (Beerenwinkel et al., 2012): (i) position-wise, by identifying single-nucleotide variants (SNV), (ii) at a local scale, by identifying patterns of SNVs that co-occur at a distance smaller than the average read length, and (iii) at a global scale, by phasing mutations over distances longer than the read length (cf. Fig. 2).

In the following, we describe computational methods for studying viral genetic diversity in accordance with the classification scheme introduced above and focusing on most recent developments.

4.1. Detecting single-nucleotide variants in virus populations

A major challenge in variant calling concerns the detection of rare SNVs. In principle, detecting low-frequency mutations is possible due to high coverages that can be attained using HTS platforms. However, the separation of genetic variants from technical noise constitutes one of the main challenges of the data analysis. While conservative thresholds for calling variants limit the sensitivity of the SNV caller, calling all variants results in poor precision. Hence, both high sensitivity and precision are desirable features for detecting low-frequency mutations.

Various statistical models accounting for sequencing errors have been proposed to boost sensitivity limits. The number of errors at each genomic position has been modeled using Poisson (Yang et al., 2013; Wang et al., 2007; Wilm et al., 2012), binomial (Macalalad et al., 2012) and beta-binomial distributions (Flaherty et al., 2012; Gerstung et al., 2012; McElroy et al., 2013). Provided reads are mapped onto a reference sequence, an SNV is called as such when a non-reference base is observed more often than expected under a given error model. However, imperfect amplification, as well as other unknown biases, usually results in a higher-than-expected variance of the nucleotide counts, an effect known as overdispersion. Among the proposed methods, beta-binomial models are able to capture overdispersion (Gerstung et al., 2012). By contrast, Poisson and binomial models do not allow for independently adjusting mean and variance, and therefore cannot account for overdispersion. For instance, the ratio between the mean and the variance in the Poisson distribution is equal to one. Since the error model determines whether the prevalence of a non-reference base is significant, overdispersion can result in systematic errors and hence should be accounted for in general.

In addition to highly sensitive methods for SNV calling, precision is also critical. This is because the number of true negatives is, in general, expected to be greater than the number of true positives. In order to reduce false positives, several tools resort to estimating position-specific error rates. Employed strategies fall into three categories: (i) methods integrating the quality scores when modeling distribution of errors (Yang et al., 2013; Wilm et al., 2012; Macalalad et al., 2012; Isakov et al., 2015), (ii) an approach using adaptive quality filters to rule out noisy base-calls (Verbist et al., 2015), and (iii) methods resorting to a control sample for estimating the background noise (Flaherty et al., 2012; Gerstung et al., 2012). In the first case, and assuming sequencing reads have been mapped to a reference sequence, non-reference bases at each locus are modeled as independent Bernoulli random variables, each of which has a distinct success probability (Yang et al., 2013; Wilm et al., 2012; Macalalad et al., 2012; Isakov et al., 2015). These probabilities are assumed to be a function of quality scores of individual bases. A tool dubbed VirVarSeq (Verbist et al., 2015) is a representative of the second category. In this case, an adaptive quality-threshold is estimated for each deep-sequenced sample. The quality threshold is determined by modeling the distribution of quality scores as a mixture of three truncated-Gaussians. Components at the lower end of the quality spectrum are interpreted as two types of errors, while the third component is interpreted as reliable calls and every nucleotide variant is treated as such. In the third case, variant counts in a heterogeneous virus populations are compared against counts in a homogeneous control sample, aiming at capturing context-specific errors (Flaherty et al., 2012; Gerstung et al., 2012). The control sample can be acquired, e.g., by sequencing monoclonal viral strains.

Further improvements in precision can be attained if systematic errors are taken into account. For instance, in the case of paired-end sequencing, there is growing evidence that sequencing errors depend on the sequencing direction and are more likely to occur on one strand than the other (Guo et al., 2012). Thus, several methods have incorporated statistical tests for strand bias (Yang et al., 2013; Wilm et al., 2012; McElroy et al., 2013).

Table 1
Comparison of pipelines for variant calling.

	ViVan	VirVarSeq
OS	Lin, Win	Unix/Lin
Language	Python	Perl, R
Dependencies	Python modules, ea-utils, SAMtools, bwa	R packages, Perl modules, Fortran compiler, bwa
Availability	No	Yes[a]
Interface	CLI/web-server	CLI
Platform	Illumina	Illumina
Input format – reads	FASTQ	FASTQ
Input format – ref.	FASTA	FASTA
Pre-processing	Quality trimming[b]	None
Alignment	Reference-based mapping	Reference-based mapping and realignment against consensus
Variant calling	Composite Bernoulli error model	Adaptive quality filtering
Applications	Coxsackie virus, Chikungunya virus	HCV
Reference	Isakov et al. (2015)	Verbist et al. (2015)

OS, operating system; Lin, Linux; Mac, Mac OS X; Win, Windows; Unix, Unix-compatible operating systems. Platform, sequencing platforms are specified if pipeline was tested on real data sets, as reported on the original publication. Input format – reads/reference; FASTA, text-based format for storing biological sequences; FASTQ, text-based format for storing biological sequences and corresponding quality scores.

[a] http://sourceforge.net/projects/virtools/?source=directory.

[b] Quality trimming is carried out by fastq-mcf, a tool from the ea-utils toolkit.

4.1.1. Analysis workflows for SNV calling

Over the last years, providing comprehensive solutions for the analysis of genomic data has become an evident necessity (Leipzig, 2016). To this end, SNV callers have been integrated into bioinformatics pipelines. Pipelines such as Viral Variance Analysis (ViVan) (Isakov et al., 2015) and VirVarSeq (Verbist et al., 2015) facilitate the characterization of the genetic diversity of virus populations, delivering SNVs from raw sequences. These pipelines combine several processing steps, including quality assessment, read alignment and variant calling (cf. Table 1), as well as downstream analyses to improve interpretability of the results. ViVan, e.g., provides several metrics and statistics on population diversity, transitions and transversion biases, synonymous and non-synonymous mutations, and gene-by-gene statistics (Isakov et al., 2015).

An overview of two surveyed pipelines is given in Table 1. We have listed some features, which include programming languages, dependencies, supported sequencing platforms, input formats and analysis steps. It can be seen that these tools are similar in many ways. For instance, both pipelines are tailored to Illumina reads, in the sense that quality scores are incorporated under the interpretation provided by Illumina platforms. Other aspects, such as dependencies on third-party software and being command line tools, make it necessary for the end-user to have at least intermediate computer skills.

4.2. Local diversity estimation

The linkage information between loci is lost when calling variants at individual genetic sites. One way to detect linkage between nucleotide variants is by identifying statistically significant patterns of co-variation in the sequencing reads. Such pairs or higher-order patterns of mutations are often referred to as *phased* sites. Phasing nucleotide variants involves detecting mutations which are observed together on multiple sites and occur more

often than expected by chance. Indeed, genetic relationships among multiple sites have been exploited in software packages such as V-phaser (Macalalad et al., 2012) and its extension V-phaser2 (Yang et al., 2013), VirVarSeq (Verbist et al., 2015), ViVaMBC (Verbist et al., 2015), CoVaMa (Routh et al., 2015) and ShoRAH (McElroy et al., 2013). An additional advantage of considering multiple sites simultaneously is that the detection limit (i.e., the minimum frequency at which variants are detectable) can be lowered below the technical noise level (McElroy et al., 2013), with a concomitant increase in statistical power.

Software packages exploiting linkage information can be subdivided into three categories (cf. Table 2). First, methods that have been tailored to performing variant calling at the codon level (Verbist et al., 2015, 2015). Software ViVaMBC (Virus Variant Model-Based Clustering) is a representative of this category. It adopts a probabilistic approach for read clustering in windows consisting of triplets of nucleotides. Underlying viral strains are modeled as the components of a multinomial mixture model. A limitation of this method is that an upper bound on the number of variants should be specified *a priori*. Methods falling in the second category include V-phaser (Macalalad et al., 2012; Yang et al., 2013) and CoVaMa (Routh et al., 2015). These methods are not limited to adjacent positions and are tailored to identifying pairs of co-occurring variants. In order to estimate a detection threshold for each pair of loci, V-phaser models the number of mismatches at both sites by constructing a composite model of independent, but not identically distributed, Bernoulli random variables (Macalalad et al., 2012). On the other hand, software CoVaMa (Co-Variation Mapper) constructs contingency tables for every pair of loci and every pair of variants, in order to compute the linkage disequilibrium (LD). A 3σ-cutoff rule is employed for assessing significance of LD values.

The third category includes methods in which the local diversity estimation is further extended to windows of the reference genome spanned by individual reads. The goal here is to phase all variant sites within such genomic regions. At this scale, local haplotype reconstruction can be regarded as a clustering problem. A basic scheme includes: (i) clustering reads based on pairwise similarities, (ii) identifying the cluster centers as predicted haplotypes, and (iii) using the cluster sizes as estimates of the haplotype frequencies.

The software ShoRAH (Short Read Assembly into Haplotypes) is a representative of the third category (Zagordi et al., 2010). ShoRAH implements local diversity estimation, coupled to SNV calling (McElroy et al., 2013), as well as global haplotype reconstruction (Eriksson et al., 2008). Local haplotype reconstruction is formulated as a probabilistic clustering approach performed in a Bayesian fashion. In a traditional clustering problem, the number of component should be specified beforehand. However, the number of underlying viral strains is, in general, unknown. Hence, and in order to capture this uncertainty, a Dirichlet process is employed as a prior probability distribution. Assignment of sequencing reads to clusters is performed iteratively on the basis of sequence similarity. In every iteration, sequencing reads are assigned with a certain probability to either an existing cluster or a new cluster. In this way, the number of components can be inferred from the data, instead of fixing it *a priori*. The centroids of read clusters are the locally reconstructed haplotypes. These predicted haplotypes are used to correct errors within read clusters. Error correction is conducted as a previous step to global haplotype reconstruction (cf. Section 4.3.1).

Other software packages have resorted to local haplotype reconstruction as the starting point for global haplotype inference (Jayasundara et al., 2015; Prosperi and Salemi, 2012; Töpfer et al., 2013; Prabhakaran et al., 2014), and are described in the next section.

Local haplotype reconstruction can be sufficient for some applications where the focus is on a genomic region which can be fully covered by individual reads. For instance, in HIV-1 infection, it is particularly relevant to study emergence of drug-resistance mutations in genes whose protein products are targeted by drugs. One such gene is the viral protease gene, which is only 297 nt long.

Some of the variant callers have been devised for either 454 or Illumina sequencing reads. Among the differences between these two sequencing platforms, the different types of predominant errors, interpretation of the quality scores, read length and throughput can be pointed out. The latter point concerning sequencing coverage is relevant for the scalability of these tools. For instance, V-phaser and ShoRAH (Prabhakaran et al., 2014), originally tested on 454 sequencing reads, have been observed to scale poorly when the coverage is on the order of tens of thousands reads. In addition, methods specialized for 454 sequencing data do not make use of the information provided by paired-end reads. The advantage of using pairing information is that distance between phased sites can be extended to longer stretches constrained only by the insert size.

4.3. Global haplotype reconstruction

Methods for local diversity estimation are limited by the length of the sequenced fragments. Correlated pairs or higher-order patterns of mutations cannot be linked at distances longer than the average read length. On the other hand, the aim in global haplotype reconstruction is to infer the genetic sequences and frequencies of the underlying viral strains over a genomic region of interest (e.g., a single gene) or across the entire genome. In either case, the size of the genomic region exceeds the read length.

Over the last decade, HTS platforms have been optimized either in terms of increased throughput and read length, or decreased error rates. Along with technological developments, several methods have been proposed to efficiently solve the global reconstruction problem from relatively short and error-prone reads. Many of these methods were originally devised for handling 454/Roche sequencing reads (Astrovskaya et al., 2011; Prosperi and Salemi, 2012; Westbrooks et al., 2008; Jojic et al., 2008; Zagordi et al., 2011), as it was the first widely-used HTS platform. Owing to the better cost-effectiveness and higher coverage offered by Illumina sequencing platforms, the focus shifted towards this technology in recent years (Mangul et al., 2014; Jayasundara et al., 2015; Töpfer et al., 2014). As the number of reads and sequencing coverage increased, more efficient algorithms were required to meet the sequencing throughput. More recently, so-called third generation sequencing platforms are gradually becoming the method of choice (Dilernia et al., 2015; Artyomenko et al., 2016; Quick et al., 2016), as latest technologies offer read lengths of tens of kilobases (kb). These developments are being driven by Pacific Biosciences (PacBio) (Eid et al., 2009) and Oxford Nanopore (Schneider and Dekker, 2012).

An orthogonal classification of the computational methods for viral haplotypes reconstruction from HTS reads has been proposed by Beerenwinkel et al. (2012). Algorithms proposed until 2012 were divided into read-graph based methods (cf. Section 4.3.1), probabilistic models (cf. Section 4.3.2) and *de novo* reconstruction (cf. Section 4.3.3). A new category is introduced here, namely haplotype reconstruction using long sequencing reads. Since methods designed for the analysis of long sequencing reads are based on hierarchical clustering, we have named them as such (cf. Section 4.3.4).

Methods based on the read graph, probabilistic models, and algorithms based on hierarchical clustering rely on the alignment of sequencing reads for the positioning and orientation of the reads with respect to a reference sequence. On the other hand, *de novo* quasispecies reconstruction methods do not rely on the existence of a reference genome and haplotypes are reconstructed directly from the sequencing reads. In the former case, a reference sequence

Table 2
Methods for local diversity estimation.

Software	Platform	Category	Approach	Output	Applications	Avail.	Ref.
VirVarSeq	Illumina	Codon-based calling	Adaptive quality filtering	Codon variants	HCV, HIV	Yes	Verbist et al. (2015)
ViVaMBC	454, Illumina	Codon-based calling	Probabilistic clustering	Codon variants	HCV (NS3)	Yes	Verbist et al. (2015)
V-Phaser/V-Phaser2	454, Illumina	Co-occurrence of pairs of variants	Composite Bernoulli error model	Variant pairs	HIV-1, WNV	Yes[a]	Macalalad et al. (2012) and Yang et al. (2013)
CoVaMa	Illlumina	Co-occurrence of pairs of variants	Linkage disequilibrium	Variant pairs	HIV (*prot*)	Yes	Routh et al. (2015)
ShoRAH	454, Illumina[b]	Local windows	Probabilistic clustering	Local haplotypes, SNVs	HIV (*pol*), HCV	Yes	McElroy et al. (2013) and Zagordi et al. (2010)

Sequencing platforms are specified if software was tested on real data sets, as reported on the original publication. Avail., availability. Ref., references.
[a] Registration required.
[b] Test for strand bias in SNV calling.

could have been previously obtained, e.g., by sequencing the sample via Sanger sequencing, or assembling the reads *de novo* into a single consensus sequence (Henn et al., 2012; Mangul et al., 2014; Jayasundara et al., 2015).

In the following, we describe in more detail different strategies for viral haplotype reconstruction. Lastly, we discuss challenges in choosing a haplotype reconstruction tool for studying diversity in mixed viral samples (cf. Section 4.3.5).

4.3.1. *Read-graph based methods for haplotype reconstruction*

The general workflow of methods based on the read graph includes mapping of sequencing reads, error correction, haplotype reconstruction and haplotype frequency estimation (cf. Fig. 3). The set of mapped reads, possibly error-corrected reads, is used to build a graph with the aim of identifying a set of paths as the viral haplotypes.

The read graph is a directed graph with vertices corresponding to non-redundant reads and edges connecting reads that agree on their non-empty overlap (Eriksson et al., 2008). A read is non-redundant if it is not fully contained within any other read. Furthermore, overlapping positions between pairs of reads and directionality of the edges are determined by the read alignment (cf. Fig. 4). A similar formulation of the read graph was independently proposed by Westbrooks et al. (2008), in which all sequencing reads are included as nodes. In this case, a more compact graph is obtained by computing the minimum transitive reduction of the

Table 3
Software packages for haplotype reconstruction based on the read graph.

Software	Sequencing mode	Error handling	Haplotype reconstruction	Haplotype frequency estimation	Avail.	Ref.
ShoRAH	Shotgun and amplicon-based	Probabilistic clustering	Minimal path cover	EM	Yes	Zagordi et al. (2011)
ViSpA	Shotgun	Binomial error model	Max-bandwidth paths	EM	Yes	Astrovskaya et al. (2011)
ShotMCF	Shotgun	Probabilistic assignment of reads to candidate haplotypes	NA[a]	Normalized flow	Yes	Skums et al. (2013)
VirA (AmpMCF)	Amplicon-based	Error-corrected reads	Multi-commodity flows	Normalized flow	Yes	Skums et al. (2013)
BIOA	Amplicon-based	Error-corrected reads[b]	Max-bandwidth paths	Frequency balancing in forked nodes	Yes	Mancuso et al. (2012)
QuRe	Amplicon-based	Poisson error model	Distribution matching	Haplotypes in decreasing order of abundance	Yes	Prosperi and Salemi (2012)
ViQuaS	PE[c]	Mutation calling	Distribution matching	Minimum frequency of constituent vertices	Yes	Jayasundara et al. (2015)
HaploClique	PE[c]	Probabilistic sequence similarity criterion	Iteratively merging max-cliques	Normalized read counts	Yes	Töpfer et al. (2014)
QColors	PE[c]	Error-corrected reads	Minimum vertex coloring	NI[d]	No	Huang et al. (2011)
VGA	PE[c]	High-fidelity sequencing protocol	Minimum vertex coloring	EM	Yes	Mangul et al. (2014)

Avail., availability. Ref., references.
[a] Candidate haplotypes are generated using the max-bandwidth method of software ViSpA.
[b] Software KEC (Skums et al., 2012) was used for error correction.
[c] PE, paired-end reads. If available, information from paired-end reads is taken into account.
[d] Not included.

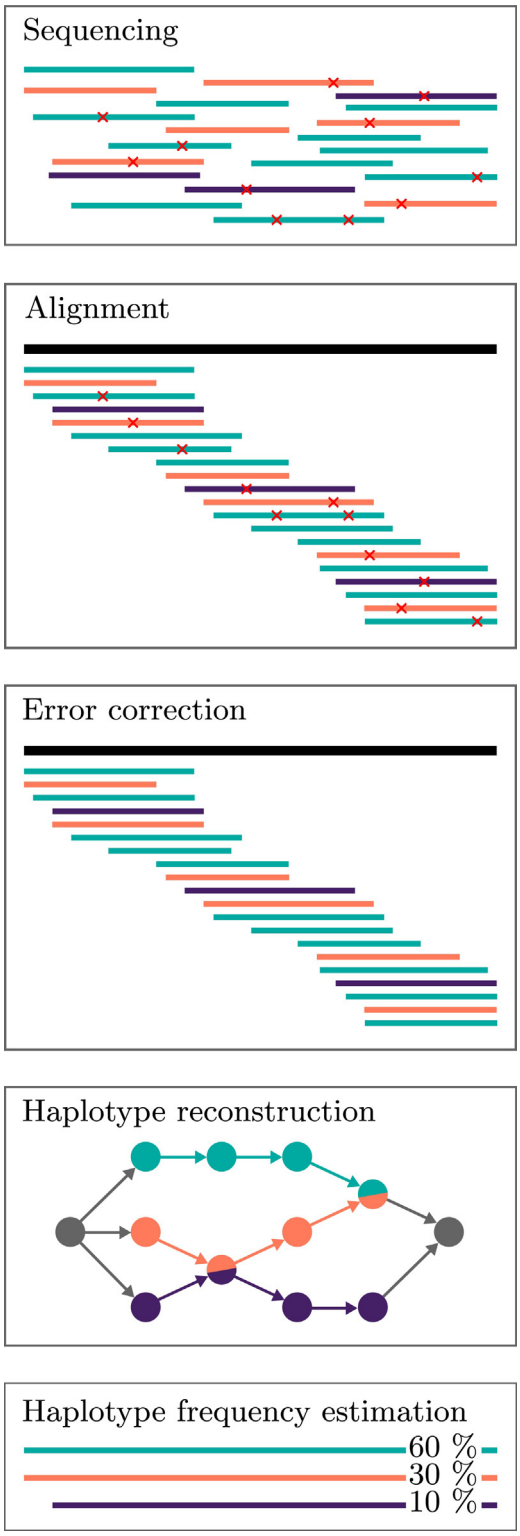

Fig. 3. Schematic workflow for global haplotype reconstruction based on the read graph. A hypothetical virus population consisting of three viral strains is deep sequenced. Reads originated from different strains are identified by distinct colors in the diagram. After sequencing, reads are aligned against a reference genome (black). Typically, aligned reads are corrected for errors, depicted here as red crosses. Corrected reads are used for building the read graph and candidate haplotypes are reconstructed as paths in the read graph. Nodes with various colors indicate that certain regions are shared among different viral strains. Finally, the relative frequencies of the reconstructed haplotypes are estimated.

read graph (Westbrooks et al., 2008). The idea here is to maximally reduce the number of edges while maintaining the existence of paths.

Source and sink nodes are typically added to the read graph (cf. gray nodes in Fig. 3). A source node is connected to all read vertices with no parents and a sink node is connected to all read vertices with no children. Thereby, any path from source to sink corresponds to a plausible haplotype. Since not every possible path corresponds to a true haplotype, and finding all possible paths leads to overestimation of the diversity of the population, the task is to find an optimal set of paths corresponding to likely viral haplotypes (cf. Fig. 4c). Several frameworks have been proposed for reconstructing viral haplotypes using the read graph, e.g., by formulating the problem as a minimal path cover problem (Eriksson et al., 2008; Zagordi et al., 2011), as a network flow problem (Westbrooks et al., 2008; Skums et al., 2013), as a maximum-bandwidth path problem (Astrovskaya et al., 2011; Mancuso et al., 2011) or using maximal clique enumeration (Töpfer et al., 2014) (cf. Table 3). In the latter case, the optimization problem is not formulated as finding paths in the read graph, but rather iteratively merging fully connected clusters of read nodes, i.e., maximal cliques, in the read graph into haplotypes of increasing length.

Recent approaches for haplotype reconstruction include methods which have been tailored to Illumina reads, such as VGA (Viral Genome Assembler) (Mangul et al., 2014), HaploClique (Töpfer et al., 2014) and ViQuaS (Jayasundara et al., 2015, 2015). Some of these methods were built upon previous algorithmic ideas, but adjusted to handle larger volumes of input reads typically produced by Illumina platforms (cf. Supplementary Table S2). VGA is based on the conflict graph introduced by Huang et al. (2011), whereas ViQuaS uses the combinatorial approach for hapolotype reconstruction proposed by Prosperi et al. (2011). Other recent developments include methods reformulating the network flow optimization problem (Astrovskaya et al., 2011; Westbrooks et al., 2008; Mancuso et al., 2012) as a multi-commodity flow approach (Skums et al., 2013). Hereafter, we explain these recent methods for haplotype reconstruction in more detail. For a comprehensive review of previously available tools, we refer to Beerenwinkel et al. (2012), as well as to Table 3 which summarizes general aspects concerning read-graph based methods.

The software ShotMCF (Skums et al., 2013) is an extension of the Viral Spectrum Assembly (ViSpA) pipeline (Astrovskaya et al., 2011) for the estimation of haplotype frequencies. Haplotype frequencies are estimated solving a network flow problem with multiple commodities, i.e., flow demands. Each commodity corresponds to a candidate haplotype generated by ViSpA and the flow through a vertex is proportional to corresponding haplotype frequencies. Particularly, and in order to account for technical errors, flow variables are weighted by the probability that corresponding reads originate from a given candidate haplotype.

ShotMCF has been designed for shotgun HTS reads. A similar approach, using multi-commodity flows, has been proposed for haplotype reconstruction using amplicon-based sequencing (Skums et al., 2013). Under sequencing protocols based on amplicons, reads are produced from pre-defined windows of a reference sequence. By design, the starting and ending positions of the amplicons with respect to a reference genome are known, as well as the overlap between amplicons. AmpMCF exploits this block structure of the amplicons for constructing the read graph. In this framework, the objective is to find a set of paths in the read graph which collectively cover all reads while minimizing the total flow. The main limitation of this method is that the number of commodities, i.e., haplotypes, needs to be specified in advance. This method has been integrated into the Viral Quasispecies Assembler pipeline (VirA).

Another analysis pipeline for viral quasispecies reconstruction has been implemented in the software ViQuaS. This software uses

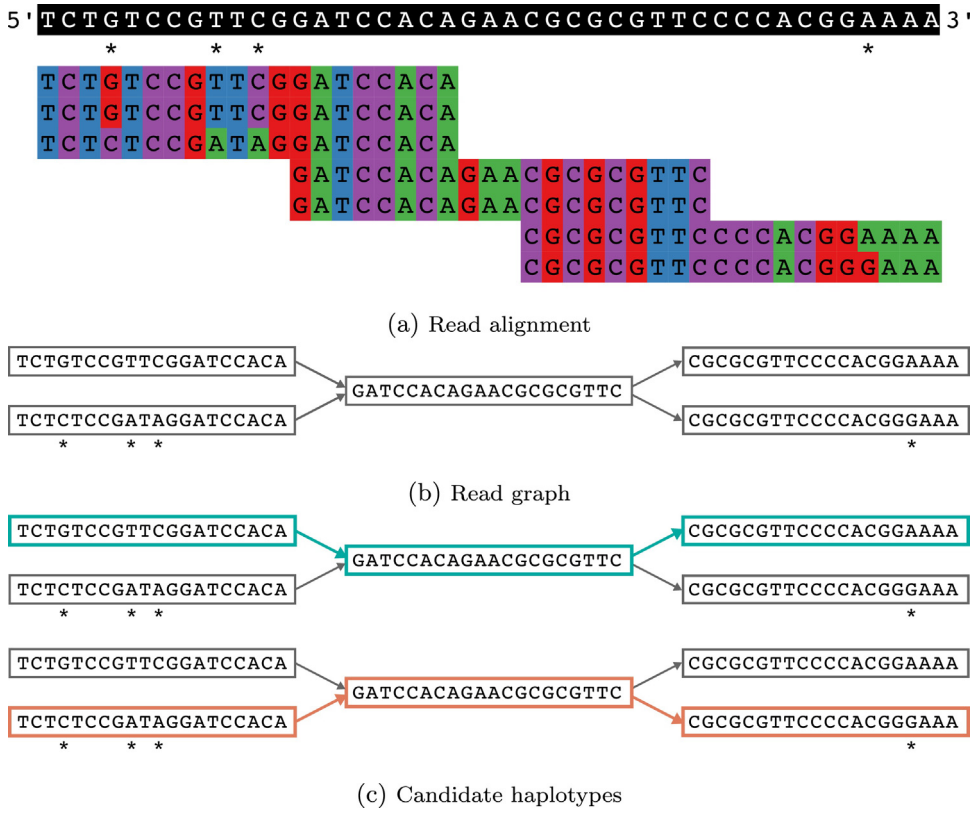

Fig. 4. Building the read graph. (a) In this example, a genomic region of length 43 bp is covered by seven reads, each of length 20 bp. Sequencing reads are aligned against a reference sequence (black) and four segregating loci are identified (asterisks). (b) From the real alignment, the read graph is constructed based on five non-redundant reads; segregating sites are indicated by asterisks. (c) Two candidate haplotypes covering all reads can be proposed from the graph and one possible solution is shown as highlighted paths in cyan and orange. In this example, closest segregating sites are further apart than a region that can be cover by any read. Therefore, there is no direct evidence of whether these SNVs occur in the same viral strain.

a reference-assisted *de novo* assembly strategy for reconstruction of haplotypes, which consists of three steps. First, reads are aligned against a reference genome and, subsequently, divided into two groups depending on whether or not each read has been perfectly aligned to the reference. Reads which did not aligned perfectly to the reference genome are assembled into contigs using the *de novo* assembly software SSAKE (Warren et al., 2007). In the second step, contigs are aligned against the reference and SNVs are naïvely detected. It may happen that reads originating from different strains have a sufficient overlap, such that they are assembled into the same contig. Therefore, in the last step, chimeric errors are corrected on the basis of reads supporting co-occurring SNVs. If there is no evidence of linking mutations, the contigs are partitioned accordingly. For the global reconstruction, contigs are combined into global haplotypes using a variation of the combinatorial approach proposed by Prosperi et al. (2011). On a benchmark study, it was observed that the pipeline reconstructed a high number of false positives, which was attributed to false *in silico* recombinants (Jayasundara et al., 2015), i.e., haplotypes which have been wrongly reconstructed as the composition of different viral strains. In order to improve on precision, a probabilisitic approach has been proposed for estimating the number of underlying strains in the virus population (Jayasundara et al., 2015). This estimate is then used as a threshold value for the number of candidate paths constructed from the read graph.

A more recent software based on the read graph is HaploClique (Töpfer et al., 2014). In this framework, the read graph is built with slight modifications. First, in case of paired-end reads, each node will correspond to a read pair. Second, in addition to sufficient overlap between two reads (or read pairs), an edge is drawn between two nodes if the corresponding reads are likely to stem from the same viral strain. The chance that two reads originate from the same haplotype is evaluated based on two criteria: (i) sequence similarity in the presence of technical errors, and (ii) compatibility of the insert sizes. The insert size criterion allows to identify structural variants. Relatively long indels are detected based on deviations from the expected insert size for read pairs in a clique. Therefore, if there are indications that the virus population poses such structural variants, HaploClique would be a suitable choice of software. Viral haplotypes are reconstructed iteratively by finding max-cliques and merging them into super-reads. A super-read is the consensus sequence of all reads in a max-clique. This iterative scheme is used to extend locally reconstructed haplotypes into full-length haplotypes, provided that the degree of genetic diversity of the virus population is sufficiently high. If there is no evidence on how to extend a super-read, the algorithm terminates. A possible downside of this method is that the run time is exponential in the read coverage. Nevertheless, HaploClique was observed to outperform other methods in terms of run time.

A complementary variant of the read-graph, the conflict graph, was initially proposed and implemented in the software QColors (Huang et al., 2011). In the conflict graph, the vertices represent reads (or read pairs) and edges are drawn between conflicting pairs of vertices, i.e., edges connect reads that do not agree on their overlap. Reconstructing viral haplotypes has been addressed in a parsimonious fashion by finding the minimum number of maximally independent sets of non-conflicting reads. This problem is equivalent to finding the minimum number of labels or colors required for coloring the vertices of the conflict graph, such that no edge connects vertices with identical colors (Huang et al.,

2011). More recently, this problem has been re-formulated as a Max-Cut problem and it has been solved using a top-down approach. The conflict graph is recursively partitioned according to its maximum cut, until components become independent (Mangul et al., 2014). Relative frequencies of the haplotypes are estimated using an expectation-maximization (EM) algorithm in a similar fashion as in the ShoRAH model (Eriksson et al., 2008). Reads are assumed to be sampled from the underlying distribution of viral strains, but here a prior probability for the viral haplotypes is used. Thus, instead of maximizing the likelihood, the posterior probability is maximized. This computational method has been implemented in the software VGA, which stands for Viral Genome Assembler (Mangul et al., 2014).

As mentioned earlier in this section, error correction can be incorporated as a step before building the read graph. Such is the case for the software ShoRAH (Zagordi et al., 2010, 2010, 2011) and QuRe (Prosperi and Salemi, 2012). Other methods, such as AmpMCF, QColors and VGA, assume reads have been corrected for technical errors, either using computational methods for error correction (Zagordi et al., 2010; Skums et al., 2012; Heo et al., 2014) or sequencing protocols based on primerIDs (Kinde et al., 2011; Lou et al., 2013; Seifert et al., 2016) (cf. Table 3). Alternatively, technical errors can be treated in a probabilistic fashion as is done in software ViSpA (Astrovskaya et al., 2011) and HaploClique (Töpfer et al., 2014).

4.3.2. *Probabilistic methods for haplotype reconstruction*

In many approaches described in the previous section, the read graph is constructed from sequencing reads which have been previously corrected for sequencing errors. Instead, the stochastic process underlying the generation of sequencing reads can be modeled explicitly. Methods falling in this category have been formulated as probabilistic models of either the sequencing process (Jojic et al., 2008) or together with a generative process for the viral haplotypes (Töpfer et al., 2013; Prabhakaran et al., 2014). The former approach has been proposed by Jojic et al. (2008), whereas the latter have been implemented in software packages PredictHaplo (Prabhakaran et al., 2014) and QuasiRecomb (Töpfer et al., 2013) (cf. Table 4).

In the model proposed by Jojic et al. (2008), observed reads are assumed to constitute a sample from a fixed number of viral strains, but a noisy sample as reads are subject to technical noise. Nevertheless, the number of underlying viral strains, which is generally unknown, needs to be specified beforehand. To circumvent this limitation, PredictHaplo and QuasiRecomb implement model selection strategies that find an optimal trade-off between sensitivity of inferring haplotypes and depth of the data.

In PredictHaplo, viral haplotypes are modeled as the components of a multinomial mixture model, in which the mixing coefficients correspond to the haplotype frequencies. Multinomial distributions are employed to capture the genetic diversity at each locus by means of haplotype- and position-specific probability tables. Moreover, and in order to avoid specification of the number of components a *priori*, a non-parametric Dirichlet process is employed as prior probability distribution, and haplotype inference is performed in a fashion similar to the local reconstruction module of ShoRAH. Full-length haplotypes are reconstructed iteratively by extending locally reconstructed haplotypes, starting from the window of aligned reads with highest coverage. Cluster assignment probabilities extracted from the locally reconstructed haplotypes are used as prior information to gradually extend the local window, and this process is carried out until the window spans the entire length of the genome or genetic region of interest.

A third generative model has been proposed based on hidden Markov models (HMM) and has been implemented in the software package QuasiRecomb (Töpfer et al., 2013). In addition to modeling mutations by position-specific probability tables, QuasiRecomb models recombination events explicitly. This feature is key when studying some RNA viruses, such as HIV, where recombination is an important source of genetic heterogeneity. As in PredictHaplo, global haplotype reconstruction is seeded on locally reconstructed haplotypes, and solved using a hierarchical assembly strategy (Di Giallonardo et al., 2014).

Both QuasiRecomb and PredictHaplo were initially tested on simulated reads mirroring 454 error patterns and read lengths. The ability of these tools to reconstruct full-length haplotypes was later validated experimentally using reads from different sequencing platforms, namely 454/Roche, Illumina and PacBio (Di Giallonardo et al., 2014).

4.3.3. De novo *assembly of viral haplotypes*

As mentioned earlier (cf. Section 3), biases induced by the read mapping hinder the reconstruction of viral haplotypes. Methods for *de novo* quasispecies assembly represent an alternative to reference-based haplotype reconstruction. To date, two reference-free methodologies, dubbed Mutant-Bin (Prabhakara et al., 2013) and MLEHaplo (Malhotra et al., 2016), have been proposed (cf. Table 5). It is worth emphasizing that assembling a single consensus genome *de novo* is not equivalent to assembling an unknown number of closely related viral haplotypes. Therefore, generic *de novo* assemblers are not well suited for the viral quasispecies reconstruction task.

A computational framework for estimating the number of viral haplotypes and their frequencies has been implemented in the method Mutant-Bin (Prabhakara et al., 2013) and later refined by Malhotra et al. (2013). This framework is based on the Lander–Waterman model of sequencing, in which reads are assumed to follow a Poisson distribution parameterized by the sequencing coverage. As such, frequencies of k-mers (i.e., substrings of length k) extracted from sequencing reads are modeled as a mixture of Poisson distributions. Expected values of the Poisson distributions correspond to the so-called *composite* frequencies, i.e., frequencies which are observed as the sum of the abundances of the underlying haplotypes sharing a given k-mer. The goal is to infer the frequencies of the underlying haplotypes, which are denoted as *basic* frequencies. A greedy strategy is used for finding a minimal set of basic frequencies explaining composite frequencies. It involves traversing the list of Poisson means in increasing order. In each iteration, an element is regarded as the frequency of an underlying haplotype if it cannot be obtained by adding basic frequencies already present in the solution set. Limitations of this method include the lack of an error model, the dependence on a uniform coverage and the assumption that different viral strains are present in the population with distinct frequencies. More importantly, genomic sequences of viral strains are not reconstructed. On the other hand, an advantage is that it allows to infer, with high precision and recall, the structure of the population when the genetic diversity is low (Malhotra et al., 2013). This is a notable advantage, because the reconstruction of viral haplotypes using any other approach becomes harder as the diversity of the sample decreases (Jayasundara et al., 2015; Eriksson et al., 2008).

More recently, a *de novo* assembly algorithm based on the *de Bruijn* graph has been proposed for estimating viral haplotypes from paired-end reads (Malhotra et al., 2016). The *de Bruijn* graph is constructed in a similar fashion to the read graph. The vertices of the graph correspond to k-mers generated from error-corrected reads and the edges connect overlapping k-mers. However, the orientation of the reads is unknown. Therefore, k-mers from the reads as well as from their reverse complements are represented in the graph. In this framework, viral haplotype reconstruction is divided into two phases. In the first phase, a fixed number of top-scoring

Table 4
Probabilistic models for haplotype reconstruction.

Method	Sequencing mode	Approach	Model selection	Inference	Avail.	Ref.
Jojic et al., 2008	Shotgun[a]	Probabilistic model for generation of sequencing reads	None	EM	No	Jojic et al. (2008)
PredictHaplo	PE[b]	Multinomial mixture model	Dirichlet process as prior probability distribution	MCMC	Yes[c]	Prabhakaran et al. (2014)
QuasiRecomb	PE[b]	Jumping HMM	BIC	EM	Yes[d]	Töpfer et al. (2013)

Inference, latent variables and underlying probability distributions are estimated from the data by maximum likelihood estimation either using the expectation-maximization algorithm (EM) or Markov chain Monte Carlo (MCMC). Avail, availability. Ref., references.
[a] Initially tested on 454 reads.
[b] PE, paired-end reads. If available the model incorporates information from paired-end reads.
[c] http://bmda.cs.unibas.ch/HivHaploTyper/.
[d] https://github.com/cbg-ethz/QuasiRecomb.

Table 5
Other methods for haplotype reconstruction.

Method	Platform/sequencing mode	Approach	Error handling	Avail.	Ref.
Mutant-Bin[a]	454/Shotgun	*De novo*	Thresholding of low-frequent *k*-mers	No	Prabhakara et al. (2013)
MLEHaplo	Illumina/PE[b]	*De novo*	Error-corrected reads[c]	Yes[d]	Malhotra et al. (2016)
Dilernia et al., 2015	PacBio	Hierarchical clustering	Binomial error model	No	Dilernia et al. (2015)
2SNV	PacBio	Hierarchical clustering	Binomial error model	Yes[e]	Artyomenko et al. (2016)

Sequencing platforms are specified if software was tested on real data sets, as reported on the original publication.
Avail., availability. Ref., references.
[a] Only applicable for haplotype frequency estimation.
[b] PE, paired-end reads. Paired-end information explicitly taken into account.
[c] Software BLESS (Heo et al., 2014) was used for error correction.
[d] https://github.com/raunaq-m/MLEHaplo.
[e] http://alan.cs.gsu.edu/NGS/?q=content/2snv_supplement.

paths per vertex are generated using a heuristic algorithm (named ViPRA). The score of a path is based on the number of read pairs covered by such path as well as on the compatibility of their insert sizes. In the second phase, the set of candidate paths is refined via backward elimination. Paths are iteratively removed until the likelihood of the remaining paths starts to decrease. This approach has been implemented in the software MLEHaplo (Malhotra et al., 2016).

4.3.4. Hierarchical clustering of long reads for reconstruction viral haplotypes

Recovering the structure of a virus population from short reads (i.e., reads shorter than the viral genome or genomic region of interest) is further hampered by the existence of relatively long regions, common to many viral strains. This is because conserved regions longer than the read length introduce ambiguities in the reconstruction of full-length viral genomes, i.e., there exists more than one plausible way to connect the relatively short reads (cf. Fig. 4). In order to bridge conserved regions, some algorithms either exploit the linkage information provided by paired-end reads (Töpfer et al., 2014) or use the relative frequencies as evidence to resolve ambiguities (Mangul et al., 2014; Prosperi and Salemi, 2012; Jojic et al., 2008; Prosperi et al., 2011). However, these strategies have severe limitations. In the former case, pairing information is limited by the length of the insert size and relies on the location of at least one of the pairs on a heterogeneous region. In the latter case, amplification and sequencing biases can lead to a non-uniform sample, resulting in deviations from the true underlying frequencies of the viral strains. This issue can be circumvented by using Pacific Bioscience or Oxford Nanopore technologies which nowadays offer read lengths that are comparable to the size of the genome of many RNA viruses.

Two methods using hierarchical clustering of long reads have been proposed for elucidating the structure of virus populations (Dilernia et al., 2015; Artyomenko et al., 2016) (cf. Table 5). Using a top down approach, reads are recursively partitioned into clusters on the basis of common SNVs, until there are no groups containing conflicting SNVs. Resulting clusters are assumed to correspond to viral strains and, consequently, haplotypes are reconstructed as the genetic consensus of each cluster. The computational workflow proposed by Dilernia et al. (2015) employs a binomial error model for calling SNVs and the pairwise distance between reads is computed as the percentage of SNVs in which the reads differ. In the framework proposed by Artyomenko et al. (2016), called 2SNV, errors are also assumed to follow a binomial distribution, but 2SNV uses linkage information between SNVs. The main limitation of the 2SNV model is reliance on the existence of linkage disequilibrium. Reads are recursively partitioned into clusters when a read cluster exhibits at least two significant segregating loci with respect to another cluster. Nonetheless, identifying viral mutant strains that differ in a single locus is a hard task (Jayasundara et al., 2015; Eriksson et al., 2008). The software 2SNV has been tested on a mixed sample obtained by error-prone PCR on the Influenza A virus PB2 segment (approx. 2 kb long). The applicability of 2SNV to longer genomic regions has not been evaluated, and it might be another limiting factor as the run time scales quadratically with respect to the number of sites evaluated.

These hierarchical clustering approaches can be regarded as local haplotype reconstruction methods, in the sense that they cannot link variants over distances larger than the read length. However, since viral genomes can be sequenced in a single run using Pacific Bioscience or Oxford Nanopore technologies, they offer the possibility to reconstruct full-length genomes. Other methods, such as the Dirichlet process mixture implemented in

ShoRAH (Zagordi et al., 2010) and PredictHaplo (Prabhakaran et al., 2014) are in principle applicable to reconstruct viral haplotypes from long reads, however, these methods have not been rigorously tested.

Other proof-of-concept applications using long reads have been also proposed to elucidate the structure of virus populations. For instance, the tag-based protocol proposed by Huang et al. (2016) relies on known mutant loci for the identification of viral strains and avoids phasing haplotypes. This methodology has been applied to analyze the linkage of HIV-1 drug resistance mutations at the haplotype level. In addition to haplotype reconstruction applications, alignment of long reads seems to be another emerging area of research. The major obstacle in aligning long reads is that due to higher error rates of, e.g., PacBio platforms and longer read lengths, previous read mappers perform poorly. Several read mappers tailored to long reads have been proposed and implemented in software packages BWA-MEM (Li, 2013), BLASR (Chaisson and Tesler, 2012), rHAT (Liu et al., 2016) and ProbAlign (Zeng et al., 2014).

4.3.5. Choice of software

When it comes to choosing a tool for analyzing a mixed sample from a virus population, the wealth of haplotype reconstruction tools is overwhelming, especially for inexperienced users. A number of review articles which were published a few years back, may prove useful when selecting a proper tool for a given application (Pandit and de Boer, 2014; Prosperi et al., 2013; Schirmer et al., 2014). However, at the time these studies were conducted, the number of tools available was scarce, and benchmarks were limited to evaluation of ShoRAH, QuRe and PredictHaplo. From these studies, the general observation is that ShoRAH and QuRe tend to over-estimate the number of haplotypes, while PredictHaplo tends to under-estimate it. Not surprisingly, PredictHaplo consistently reports the lowest number of false positives, resulting in high precision oftentimes at the cost of low recall.

Other comparative assessments have been included in publications of the latest methods. Usually, these studies emphasize how the authors' software improves over previous tools, and although subjectivity may be questionable, the main difficulty is a lack of standardization. Henceforth, we summarize the most general findings.

Factors influencing reliability of viral haplotype reconstruction include the ratio between the read length and the genome size, the depth of coverage, technical error rates, abundances of viral strains and the underlying genetic heterogeneity of the virus population. The first three factors are part of the experimental design and, therefore, can be controlled, while the latter are intrinsic to the virus population. In general, the longer the reads, the higher the coverage, the more abundant the viral strain and the larger the degree of diversity, the better one can expect any given method to perform (Mangul et al., 2014; Jayasundara et al., 2015; Malhotra et al., 2016; Pandit and de Boer, 2014; Eriksson et al., 2008; Prabhakaran et al., 2014; Töpfer et al., 2014; Skums et al., 2013; Prosperi et al., 2013; Schirmer et al., 2014; Zagordi et al., 2012). Coverage influences the minimum frequency at which a rare mutant can be identified, whereas the read length and the genetic diversity affect the capability of bridging gaps between conserved regions in different haplotypes. Additionally, most abundant haplotypes are better represented in the sample than low abundant counterparts, thus, are oftentimes reconstructed more accurately.

Although some of the aforementioned factors can be modified as part of the experimental design, there are some technical limitations when choosing a software for analyzing the data. For instance, independent studies have indicated that QuRe aborted execution, among other reasons, due to the large volume of input reads (Jayasundara et al., 2015; Prabhakaran et al., 2014; Töpfer et al., 2014; Schirmer et al., 2014). In order to ease the selection process, we have listed some aspects reported in the original publications (cf. Supplementary Table S2). These data are intended to illustrate in which ranges of, e.g., read lengths or volumes of input reads, performance of a given tool has been studied. A Haplotype reconstruction tool may well run outside these ranges.

In terms of run times, and aggregating results from different benchmarks, it has been observed that latest read-graph based algorithms, such as ViQuaS and HaploClique are faster than PredictHaplo, which in turn is faster than ShoRAH, ViSpA and QuRe (Jayasundara et al., 2015; Astrovskaya et al., 2011; Prabhakaran et al., 2014; Töpfer et al., 2014; Schirmer et al., 2014).

To the best of our knowledge, all software packages developed to date have been developed for research purposes. Unfortunately, usability, portability and maintainability are not priorities in this setting. Most, if not all, software are distributed as command line tools (cf. Supplementary Table S2) and the few available pipelines lack integration into scientific workflow systems. Therefore, basic to advanced computational skills are a pre-requisite from the user.

5. Conclusions and future directions

HTS technologies have opened up new avenues for studying genetic heterogeneity of virus populations at an unprecedented level of detail. However, since reads are error-prone and typically shorter than the targeted genomic region, HTS platforms provide an incomplete and imperfect sample of the virus population. Biases and errors introduced during library preparation steps, amplification and actual sequencing can be ameliorated using sample controls and superior experimental protocols (e.g., primerIDs, CirSeq). On the other hand, reference biases induced by the read alignment remain challenging, and current aligners leave a lot of room for improvement.

Several methods for SNV calling have been proposed and implemented in the past five years. Nowadays, it is possible to detect variants in the population at relative frequencies below 1%. Information on low-frequency mutations is of relevance for antiviral treatment and thus we foresee applications aimed at routine practice in clinical virology, substituting diagnosis based on Sanger sequencing.

Many strategies have been proposed as solutions for the viral quasispecies reconstruction problem. However, it is difficult to comparatively assess their performance. This is largely due to (i) several factors influencing the accuracy of methods for haplotype reconstruction, as well as (ii) lack of standardized metrics or (iii) validation standards. Firstly, there are many factors influencing reliability of haplotype reconstruction algorithms, including aspects related to the sequencing platform of choice (e.g., read length, coverage, error rates) and intrinsic to the virus population (e.g., genetic diversity, strain prevalence). Secondly, it would be desirable to have a widely accepted performance metric which quantifies the ability of reconstructing viral haplotypes accounting for aforementioned factors or with respect to theoretical limits. The latter can be estimated, e.g., using the Lander–Waterman model. Under the assumption of uniform coverage, the Lander–Waterman model provides a theoretical bound on the relative frequency at which viral strains can be identified (Jayasundara et al., 2015; Eriksson et al., 2008), but, in addition to several simplifying assumptions, ignores the combinatorial problem that arises while merging short reads into full-length haplotypes. Lastly, from a practical perspective, many software packages have been tested only on simulated data sets, where factors such as faithfulness of simulated error profiles and assumptions on the structure of the virus population may be questionable. Furthermore, when the performance of haplotype reconstruction tools is evaluated on real data sets, reconstruction accuracy remains an issue (Pandit and de Boer,

2014). Therefore, further validations in experimental settings are needed, e.g., by using well-characterized samples of mixed viral populations (Prosperi et al., 2013; Di Giallonardo et al., 2014). Other topics which are not commonly addressed are the algorithmic complexity and scalability of implemented methods, because heuristics are oftentimes not amenable to asymptotic complexity analysis.

The plethora of computational methods for diversity assessment, as well as unresolved challenges concerning their benchmarking, renders the choice of a proper tool for a given application a non-trivial task. A potential solution to the lack of robust benchmark standards would be to run comparison contests, as has successfully been done to comparatively assess genome assemblers or metagenomic pipelines (Marx, 2016). Another critical aspect concerns usability of tools for haplotype reconstruction. Most software packages are tailored to well-experienced bioinformaticians, limiting their applicability. The need for standardized analysis work-flows, as well as improved usability, is driving the development and implementation of bioinformatics pipeline frameworks (Leipzig, 2016). In spite of technical developments, additional commitment from the community is needed in order to embrace more flexible and easily extensible pipelines, without overlooking user-friendliness. As to the latter, e.g., bioinformatics pipelines often rely on third-party software to be pre-installed. A promising trend is to provide software packages in containers, with a minimal filesystem which includes all dependencies. In this way, complex software building steps on the part of the user are made redundant and analyses can be made reproducible.

Along with recent advances in sequencing technologies, a promising direction appears to be reconstructing viral haplotypes from long sequencing reads. As error rates and read coverage offered by so-called third-generation sequencing technologies continue to improve, we anticipate that novel algorithmic solutions for the analysis of these data will become an active area of research.

Acknowledgements

We thank Marek Pikulski for his insightful comments and critical reading.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.virusres.2016.09.016.

References

Acevedo, A., Andino, R., 2014. Library preparation for highly accurate population sequencing of RNA viruses. Nat. Protoc. 9, 1760–1769, http://dx.doi.org/10.1038/nprot.2014.118.

Archer, J., Rambaut, A., Taillon, B.E., Harrigan, P.R., Lewis, M., Robertson, D.L., 2010. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time-an ultra-deep approach. PLoS Comput. Biol. 6 (12), 1–11, http://dx.doi.org/10.1371/journal.pcbi.1001022.

Artyomenko, A., Wu, N.C., Mangul, S., Eskin, E., Sun, R., Zelikovsky, A., 2016. Long single-molecule reads can resolve the complexity of the Influenza virus composed of rare, closely related mutant variants. In: Research in Computational Molecular Biology: 20th Annual Conference, RECOMB 2016. Proceedings, Santa Monica, CA, USA, April 17–21. Springer International Publishing, pp. 164–175, http://dx.doi.org/10.1007/978-3-319-31957-5_12.

Astrovskaya, I., Tork, B., Mangul, S., Westbrooks, K., Măndoiu, I., Balfe, P., Zelikovsky, A., 2011. Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinform. 12 (6), 1–10, http://dx.doi.org/10.1186/1471-2105-12-S6-S1.

Audsley, J., Littlejohn, M., Yuen, L., Sasadeusz, J., Ayres, A., Desmond, C., Spelman, T., Lau, G., Matthews, G.V., Avihingsanon, A., Seaberg, E., Philp, F., Saulynas, M., Ruxrungtham, K., Dore, G.J., Locarnini, S.A., Thio, C.L., Lewin, S.R., Revill, P.A., 2010. HBV mutations in untreated HIV-HBV co-infection using genomic length sequencing. Virology 405 (2), 539–547, http://dx.doi.org/10.1016/j.virol.2010.06.038.

Avidor, B., Girshengorn, S., Matus, N., Talio, H., Achsanov, S., Zeldis, I., Fratty, I.S., Katchman, E., Brosh-Nissimov, T., Hassin, D., Alon, D., Bentwich, Z., Yust, I., Amit, S., Forer, R., Shultsman, I.V., Turner, D., 2013. Evaluation of a benchtop HIV ultradeep pyrosequencing drug resistance assay in the clinical laboratory. J. Clin. Microbiol. 51 (3), 880–886, http://dx.doi.org/10.1128/JCM.02652-12.

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., Song, Y.-Q., 2011. Evaluation of next-generation sequencing software in mapping and assembly. J. Hum. Genet. 56 (6), 406–414, http://dx.doi.org/10.1038/jhg.2011.43.

Beerenwinkel, N., Günthard, H.F., Roth, V., Metzner, K.J., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. 3, 329, http://dx.doi.org/10.3389/fmicb.2012.00329.

Berthet, F.-X., Zeller, H.G., Drouet, M.-T., Rauzier, J., Digoutte, J.-P., Deubel, V., 1997. Extensive nucleotide changes and deletions within the envelope glycoprotein gene of Euro-African West Nile viruses. J. Gen. Virol. 78, 2293–2297, http://dx.doi.org/10.1099/0022-1317-78-9-2293.

Bonhoeffer, S., Nowak, M.A., 1997. Pre-existence and emergence of drug resistance in HIV-1 infection. Proc. Biol. Sci. 264 (1382), 631–637, http://dx.doi.org/10.1098/rspb.1997.0089.

Borucki, M.K., Allen, J.E., Chen-Harris, H., Zemla, A., Vanier, G., Mabery, S., Torres, C., Hullinger, P., Slezak, T., 2013. The role of viral population diversity in adaptation of bovine coronavirus to new host environments. PLoS ONE 8 (1), e52752, http://dx.doi.org/10.1371/journal.pone.0052752.

Burrows, M., Wheeler, D.J., 1994. A block-sorting lossless data compression algorithm, Tech. Rep. 124. Digital Equipment Corporation, Systems Research Center.

Caboche, S., Audebert, C., Lemoine, Y., Hot, D., 2014. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. BMC Genomics 15 (1), 1–16, http://dx.doi.org/10.1186/1471-2164-15-264.

Chaisson, M.J., Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinform. 13 (1), 1–18, http://dx.doi.org/10.1186/1471-2105-13-238.

Cheval, J., Sauvage, J., Frangeul, L., Dacheux, L., Guigon, G., Dumey, N., Pariente, K., Rousseaux, C., Dorange, F., Berthet, N., Brisse, S., Moszer, I., Bourhy, H., Manuguerra, C.J., Lecuit, M., Burguiere, A., Caro, V., Eloit, M., 2011. Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples. J. Clin. Microbiol. 49 (9), 3268–3275, http://dx.doi.org/10.1128/JCM.00850-11.

Cuevas, J.M., Geller, R., Garijo, R., López-Aldeguer, J., Sanjuán, R., 2015. Extremely high mutation rate of HIV-1 in vivo. PLoS Biol. 13 (9), 1–19, http://dx.doi.org/10.1371/journal.pbio.1002251.

Di Giallonardo, F., Töpfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., Schmutz, S., Campbell, N.K., Joos, B., Lecca, M.R., Patrignani, A., Daümer, M., Beisel, C., Rusert, P., Trkola, A., Günthard, H.F., Roth, V., Beerenwinkel, N., Metzner, K.J., 2014. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucl. Acids Res. 42 (14), e115, http://dx.doi.org/10.1093/nar/gku537.

Dilernia, D.A., Chien, J.-T., Monaco, D.C., Brown, M.P., Ende, Z., Deymier, M.J., Yue, L., Paxinos, E.E., Allen, S., Tirado-Ramos, A., Hunter, E., 2015. Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. Nucl. Acids Res. 43 (20), e129, http://dx.doi.org/10.1093/nar/gkv630.

Domingo, E., Holland, J., 1997. RNA virus mutations and fitness for survival. Annu. Rev. Microbiol. 51, 151–178.

Domingo, E., Escarmí, C., Lázaro, E., Manrubia, S.C., 2005. Quasispecies dynamics and RNA virus extinction. Virus Res. 107 (2), 129–139, Virus Entry into Error Catastrophe as a New Antiviral Strategy. doi:10.1016/j.virusres.2004.11.003.

Domingo, E., 2015. Virus as Populations: Composition, Complexity, Dynamics, and Biological Implications. Academic Press.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267–276, http://dx.doi.org/10.1038/nrg2323.

Eddy, S., 2003. HMMER user's guide. biological sequence analysis using profile hidden Markov models.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-time dna sequencing from single polymerase molecules. Science 323 (5910), 133–138, http://dx.doi.org/10.1126/science.1162986.

Eigen, M., Schuster, P., 1977. The hypercycle. A principle of natural self-organization. A. Emergence of the hypercycle. Naturwissenschaften 64, 541–565.

Eigen, M., Schuster, P., 1978. The hypercycle. A principle of natural self-organization. B. The abstract hypercycle. Naturwissenschaften 65, 7–41.

Eigen, M., Schuster, P., 1978. The hypercycle. A principle of natural self-organization. C. The realistic hypercycle. Naturwissenschaften 65, 341–369.

Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.-Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R.W., Beerenwinkel, N., 2008. Viral population estimation using

pyrosequencing. PLoS Comput. Biol. 4 (5), e1000074, http://dx.doi.org/10.1371/journal.pcbi.1000074.
Ferragina, P., Manzini, G.,2000. Opportunistic data structures with applications. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science. IEEE Computer Society, pp. 390–398, http://dx.doi.org/10.1109/SFCS.2000.892127.
Flaherty, P., Natsoulis, G., Muralidharan, O., Winters, M., Buenrostro, J., ell, J., Brown, S., Holodniy, M., Zhang, N., Ji, H.P., 2012. Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucl. Acids Res. 40 (1), e2, http://dx.doi.org/10.1093/nar/gkr861.
Fonseca, N.A., Rung, J., Brazma, A., Marioni, J.C., 2012. Tools for mapping high-throughput sequencing data. Bioinformatics 28 (24), 3169–3177, http://dx.doi.org/10.1093/bioinformatics/bts605.
Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B.H., Bhattacharya, T., Korber, B., 2002. Diversity considerations in HIV-1 vaccine selection. Science 296, 2354–2360, http://dx.doi.org/10.1126/science.1070441.
Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., Beerenwinkel, N., 2012. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nat. Commun. 3, 811, http://dx.doi.org/10.1038/ncomms1814.
Gianella, S., Richman, D.D., 2010. Minority variants of drug-resistant HIV. J. Infect. Dis. 202 (5), 657–666, http://dx.doi.org/10.1086/655397.
Guglietta, S., Pantaleo, G., Graziosi, C., 2010. Long sequence duplications, repeats, and palindromes in HIV-1 gp120: length variation in V4 as the product of misalignment mechanism. Virology 399 (1), 167–175, http://dx.doi.org/10.1016/j.virol.2009.12.030.
Guo, Y., Li, J., Li, C.-I., Long, J., Samuels, D.C., Shyr, Y., 2012. The effect of strand bias in illumina short-read sequencing data. BMC Genomics 13, 666, http://dx.doi.org/10.1186/1471-2164-13-666.
Henn, M.R., Boutwell, C.L., Charlebois, P., Lennon, N.J., Power, K.A., Macalalad, A.R., Berlin, A.M., Malboeuf, C.M., Ryan, E.M., Gnerre, S., Zody, M.C., Erlich, R.L., Green, L.M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A.K., Dudek, T., Tully, D., Newman, R., Axten, K.L., Gladden, A.D., Battis, L., Kemper, M., Zeng, Q., Shea, T.P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Günthard, H.F., Brumme, Z.L., Brumme, C.J., Bazner, S., Rychert, J., Tinsley, J.P., Mayer, K.H., Rosenberg, E., Pereyra, F., Levin, J.Z., Young, S.K., Jessen, H., Marcus Altfeld, B.W., Birren, Walker, B.D., Allen, T.M., 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 8, e1002529, http://dx.doi.org/10.1371/journal.ppat.1002529.
Heo, Y., Wu, X.-L., Chen, D., Ma, J., Hwu, W.-M., 2014. BLESS: bloom filter-based error correction solution for high-throughput sequencing reads. Bioinformatics 30 (10), 1354–1362, http://dx.doi.org/10.1093/bioinformatics/btu030.
Hong, L.Z., Hong, S., Wong, H.T., Aw, P.P., Cheng, Y., Wilm, A., de Sessions, P.F., Lim, S.G., Nagarajan, N., Hibberd, M.L., Quake, S.R., Burkholder, W.F., 2014. BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads. Genome Biol. 15 (11), 1–14, http://dx.doi.org/10.1186/s13059-014-0517-9.
Huang, A., Kantor, R., DeLong, A., Schreier, L., Istrai, S., 2011. Qcolors: an algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads. In Silico Biol. 11 (5,6), 193–201, 2012, 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops.
Huang, D.W., Raley, C., Jiang, M., K, Zheng, X., Liang, D., Rehman, M.T., Dewar, R.L., 2016. Towards better precision medicine: PacBio single-molecule long reads resolve the interpretation of HIV drug resistant mutation profiles at explicit quasispecies (haplotype) level. J. Data Min. Genomics Proteomics 7 (1), 182, http://dx.doi.org/10.4172/2153-0602.1000182.
Hunt, M., Gall, A., Ong, S.H., Brener, J., Ferns, B., Goulder, P., Nastouli, E., Keane, J.A., Kellam, P., Otto, T.D., 2015. IVA: accurate *de novo* assembly of RNA virus genomes. Bioinformatics 31 (14), 2374–2376, http://dx.doi.org/10.1093/bioinformatics/btv120.
Hurwitz, B.L., Sullivan, M.B., 2013. The pacific ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. PLoS ONE 8 (2), 1–12, http://dx.doi.org/10.1371/journal.pone.0057355.
Isakov, O., Bordería, A.V., Golan, D., Hamenahem, A., Celniker, G., Yoffe, L., Blanc, H., Vignuzzi, M., Shomron, N., 2015. Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. Bioinformatics 31 (13), 2141–2150, http://dx.doi.org/10.1093/bioinformatics/btv101.
Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A., Swanstrom, R., 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer ID. Proc. Natl. Acad. Sci. U. S. A. 108 (50), 20166–20171, http://dx.doi.org/10.1073/pnas.1110064108.
Jayasundara, D., Saeed, I., Maheswararajah, S., Chang, B., Tang, S.-L., Halgamuge, S.K., 2015. ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. Bioinformatics 31 (6), 886–896, http://dx.doi.org/10.1093/bioinformatics/btu754.
Jayasundara, D., Saeed, I., Chang, B., Tang, S.-L., Halgamuge, S.K., 2015. Accurate reconstruction of viral quasispecies spectra through improved estimation of strain richness. BMC Bioinform. 16 (18), 1–12, http://dx.doi.org/10.1186/1471-2105-16-S18-S3.
Johnson, J.A., Li, J.-F., Wei, X., Lipscomb, J., Irlbeck, D., Craig, C., Smith, A., Bennett, D.E., Monsour, M., Sandstrom, P., Lanier, E.R., Heneine, W., 2008. Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naïve populations and associate with reduced treatment efficacy. PLoS Med. 5 (7), 1–11, http://dx.doi.org/10.1371/journal.pmed.0050158.
Jojic, V., Hertz, T., Jojic, N., 2008. Population sequencing using short reads: HIV as a case study. In: Proc. Pacific Symp. Biocomputing, pp. 114–125.
Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., 2011. Detection and quantification of rare mutations with massively parallel sequencing. Proc. Natl. Acad. Sci. U. S. A. 108 (23), 9530–9535, http://dx.doi.org/10.1073/pnas.1105422108.
Kuroda, M., Katano, H., Nakajima, N., Tobiume, M., Ainai, A., Sekizuka, T., Hasegawa, H., Tashiro, M., Sasaki, Y., Arakawa, Y., Hata, S., Watanabe, M., Sata, T., 2010. Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by *de novo* sequencing using a next-generation DNA sequencer. PLoS ONE 5 (4), e10256, http://dx.doi.org/10.1371/journal.pone.0010256.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359, http://dx.doi.org/10.1038/nmeth.1923.
Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (3), 1–10, http://dx.doi.org/10.1186/gb-2009-10-3-r25.
Lauring, A.S., Andino, R., 2010. Quasispecies theory and the behavior of RNA viruses. PLoS Pathog 6 (7), e1001005, http://dx.doi.org/10.1371/journal.ppat.1001005.
Lee, W.-P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., Marth, G.T., 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS ONE 9 (3), 1–11, http://dx.doi.org/10.1371/journal.pone.0090581.
Leipzig, J., 2016. A review of bioinformatic pipeline frameworks. Brief. Bioinform., 1–7, http://dx.doi.org/10.1093/bib/bbw020.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25 (14), 1754–1760, http://dx.doi.org/10.1093/bioinformatics/btp324.
Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26 (5), 589–595, http://dx.doi.org/10.1093/bioinformatics/btp698.
Li, H., Homer, N., 2010. A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 11 (5), 473–483, http://dx.doi.org/10.1093/bib/bbq015.
Li, R., Li, Y., Kristiansen, K., Wang, J., 2008. SOAP: short oligonucleotide alignment program. Bioinformatics 24 (5), 713–714, http://dx.doi.org/10.1093/bioinformatics/btn025.
Li, H., Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997.
Liu, B., Guan, D., Teng, M., Wang, Y., 2016. rHAT: fast alignment of noisy long reads with regional hashing. Bioinformatics 32 (11), 1625–1631, http://dx.doi.org/10.1093/bioinformatics/btv662.
Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., Pallen, M.J., 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30, 434–439, http://dx.doi.org/10.1038/nbt.2198.
Lou, D.I., Hussmann, J.A., McBee, R.M., Acevedo, A., Andino, R., Press, W.H., Sawyer, S.L., 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc. Natl. Acad. Sci. U. S. A. 110 (49), 19872–19877, http://dx.doi.org/10.1073/pnas.1319590110.
Lunter, G., Goodson, M., 2011. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. Genome Res. 21 (6), 936–939, http://dx.doi.org/10.1101/gr.111120.110.
Macalalad, A.R., Zody, M.C., Charlebois, P., Lennon, N.J., Newman, R.M., Malboeuf, C.M., Ryan, E.M., Boutwell, C.L., Power, K.A., Brackney, D.E., Pesko, K.N., Levin, J.Z., Ebel, G.D., Allen, T.M., Birren, B.W., Henn, M.R., 2012. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput. Biol. 8 (3), e1002417, http://dx.doi.org/10.1371/journal.pcbi.1002417.
Malhotra, R., Prabhakara, S., Poss, M., Acharya, R., 2013. Estimating viral haplotypes in a population using k-mer counting. In: Pattern Recognition in Bioinformatics: 8th IAPR International Conference, PRIB 2013. Proceedings, Nice, France, June 17–20, 2013. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 265–276, http://dx.doi.org/10.1007/978-3-642-39159-0_24.
Malhotra R., Mukhopadhyay M., Wu S., Rodrigo A., Poss M. Acharya R., Maximum likelihood *de novo* reconstruction of viral populations using paired end sequencing data. arXiv:1502.04239.
Mancuso, N., Tork, B., Skums, P., Mvandoiu, I., Zelikovsky, A., 2011. Viral quasispecies reconstruction from amplicon 454 pyrosequencing reads. In Silico Biol., 1–13.
Mancuso, N., Skums, P., Ganova-Raeva, L., Mvandoiu, I., Zelikovsky, A., 2012. Reconstructing viral quasispecies from NGS amplicon reads. In Silico Biol. 11 (5–6), 237–249, http://dx.doi.org/10.1109/BIBMW.2011.6112360.
Mangul, S., Wu, N.C., Mancuso, N., Zelikovsky, A., Sun, R., Eskin, E., 2014. Accurate viral population assembly from ultra-deep sequencing data. Bioinformatics 30 (12), i329–i337, http://dx.doi.org/10.1093/bioinformatics/btu295.
Mansky, L.M., Temin, H.M., 1995. Lower *in vivo* mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. 69 (8), 5087–5094.
Marx, V., 2016. Microbiology: the road to strain-level identification. Nat. Methods 13, 401–404, http://dx.doi.org/10.1038/nmeth.3837.
McElroy, K., Zagordi, O., Bull, R., Luciani, F., Beerenwinkel, N., 2013. Accurate single nucleotide variant detection in viral populations by combining probabilistic

clustering with a statistical test of strand bias. BMC Genomics 14 (1), 1–12, http://dx.doi.org/10.1186/1471-2164-14-501.
McElroy, K., Thomas, T., Luciani, F., 2014. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microb. Inform. Exp. 4 (1), 1–14, http://dx.doi.org/10.1186/2042-5783-4-1.
Metzner, K.J., Giulieri, S.G., Knoepfel, S.A., Rauch, P., Burgisser, P., Yerly, S., Günthard, H.F., Cavassini, M., 2009. Minority quasispecies of drug-resistant HIV-1 that lead to early failure in treatment-naïve and -adherent patients. Clin. Infect. Dis. 48 (2), 239–247, http://dx.doi.org/10.1086/595703.
Mount, D.W., 2009. Using hidden Markov models to align multiple sequences. Cold Spring Harb. Protoc. (7), http://dx.doi.org/10.1101/pdb.top41.
Nowak, M.A., Anderson, R.M., McLean, A.R., Wolfs, T.F., Goudsmit, J., May, R.M., 1991. Antigenic diversity thresholds and the development of AIDS. Science 254 (5034), 963–969, http://dx.doi.org/10.1126/science.1683006.
Nowak, M.A., 1992. What is a quasispecies? Trends Ecol. Evol. 7 (11), 118–121.
Palmer, S., Kearney, M., Maldarelli, F., Halvas, E.K., Bixby, C.J., Bazmi, H., Rock, D., Falloon, J., Davey Jr., R.T., Dewar, R.L., Metcalf, J.A., Hammer, S., Mellors, J.W., Coffin, J.M., 2005. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. J. Clin. Microbiol. 43 (1), 406–413, http://dx.doi.org/10.1128/JCM.43.1.406-413.2005.
Pandit, A., de Boer, R.J., 2014. Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants. Retrovirology 11 (1), 1–15, http://dx.doi.org/10.1186/1742-4690-11-56.
Park, S., Kim, S., Song, D., Park, B., 2014. Novel porcine epidemic diarrhea virus variant with large genomic deletion, South Korea. Emerg. Infect. Dis. 20 (12), 2089–2092, http://dx.doi.org/10.3201/eid2012.131642.
Prabhakara, S., Malhotra, R., Acharya, R., Poss, M., 2013. Mutant-Bin: unsupervised haplotype estimation of viral population diversity without reference genome. J. Comput. Biol. 20 (6), 453–463, http://dx.doi.org/10.1089/cmb.2012.0174.
Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N., Roth, V., 2014. HIV haplotype inference using a propagating Dirichlet process mixture model. IEEE/ACM Trans. Comput. Biol. Bioinform. 11 (1).
Prosperi, M.C., Salemi, M., 2012. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics 28, 132–133, http://dx.doi.org/10.1093/bioinformatics/btr627.
Prosperi, M.C.F., Prosperi, L., Bruselles, A., Abbate, I., Rozera, G., Vincenti, D., Solmone, M.C., Capobianchi, M.R., Ulivi, G., 2011. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. BMC Bioinform. 12 (1), 1–13, http://dx.doi.org/10.1186/1471-2105-12-5.
Prosperi, M.C.F., Yin, L., Nolan, D.J., Lowe, A.D., Goodenow, M.M., Salemi, M., 2013. Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. Sci. Rep. 3, 2837, http://dx.doi.org/10.1038/srep02837.
Quick, J., Loman, N.J., Duraffour, S., Simpson, J.T., Severi, E., Cowley, L., Bore, J.A., Koundouno, R., Dudas, G., Mikhail, A., Ouédraogo, N., Afrough, B., Bah, A., Baum, J.H.J., Becker-Ziaja, B., Boettcher, J.P., Cabeza-Cabrerizo, M., Camino-Sánchez, A., Carter, L.L., Doerrbecker, J., Enkirch, T., García-Dorival, I., Hetzelt, N., Hinzmann, J., Holm, T., Kafetzopoulou, L.E., Koropogui, M., Kosgey, A., Kuisma, E., Logue, C.H., Mazzarelli, A., Meisel, S., Mertens, M., Michel, J., Ngabo, D., Nitzsche, K., Pallasch, E., Patrono, L.V., Portmann, J., Repits, J.G., Rickett, N.Y., Sachse, A., Singethan, K., Vitoriano, I., Yemanaberhan, R.L., Zekeng, E.G., Racine, T., Bello, A., Sall, A.A., Faye, O., Faye, O., Magassouba, N., Williams, C.V., Amburgey, V., Winona, L., Davis, E., Gerlach, J., Washington, F., Monteil, V., Jourdain, M., Bererd, M., Camara, A., Somlare, H., Camara, A., Gerard, M., Bado, G., Baillet, B., Delaune, D., Nebie, K.Y., Diarra, A., Savane, Y., Pallawo, R.B., Jaramillo Gutierrez, G., Milhano, N., Roger, I., Williams, C.J., Yattara, F., Lewandowski, K., Taylor, J., Rachwal, P., Turner, D.J., Pollakis, G., Hiscox, J.A., Matthews, D.A., O'Shea, M.K., Johnston, A.M., Wilson, D., Hutley, E., Smit, E., Di Caro, A., Wölfel, R., Stoecker, K., Fleischmann, E., Gabriel, M., Weller, S.A., Koivogui, L., Diallo, B., Keïta, S., Rambaut, A., Formenty, P., Günther, S., Carroll, M.W., 2016. Real-time, portable genome sequencing for ebola surveillance. Nature 530 (7589), 228–232, http://dx.doi.org/10.1038/nature16996.
Reguera, J., Ordo no, D., Santiago, C., Enjuanes, L., Casasnovas, J.M., 2011. Antigenic modules in the N-terminal S1 region of the transmissible gastroenteritis virus spike protein. J. Gen. Virol. 92 (5), 1117–1126, http://dx.doi.org/10.1099/vir.0.027607-0.
Routh, A., Chang, M.W., Okulicz, J.F., Johnson, J.E., Torbett, B.E., 2015. CoVaMa: Co-Variation Mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data. Methods 91, 40–47, http://dx.doi.org/10.1016/j.ymeth.2015.09.021.
Rozera, G., Abbate, I., Vlassi, C., Giombini, E., Lionetti, R., Selleri, M., Zaccaro, P., Bartolini, B., Corpolongo, A., D'Offizi, G., Baiocchini, A., Del Nonno, F., Ippolito, G., Capobianchi, M., 2014. Quasispecies tropism and compartmentalization in gut and peripheral blood during early and chronic phases of HIV-1 infection: possible correlation with immune activation markers. Clin. Microbiol. Infect. 20 (3), O157–O166, http://dx.doi.org/10.1111/1469-0691.12367.
Schirmer, M., Sloan, W.T., Christopher, Q., 2014. Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. Brief. Bioinform. 15 (3), 431–442, http://dx.doi.org/10.1093/bib/bbs081.
Schneider, G.F., Dekker, C., 2012. DNA sequencing with nanopores. Nat. Biotechnol. 30, 326–328, http://dx.doi.org/10.1038/nbt.2181.
Seifert, D., Di Giallonardo, F., Metzner, K.J., Günthard, H.F., Beerenwinkel, N., 2015. A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory. Genetics 199 (1), 191–203, http://dx.doi.org/10.1534/genetics.114.172312.
Seifert, D., Di Giallonardo, F., Töpfer, A., Singer, J., Schmutz, S., Günthard, H.F., Beerenwinkel, N., Metzner, K.J., 2016. A comprehensive analysis of primer IDs to study heterogeneous HIV-1 populations. J. Mol. Biol. 428 (1), 238–250, http://dx.doi.org/10.1016/j.jmb.2015.12.012.
Skums, P., Dimitrova, Z., Campo, D.S., Vaughan, G., Rossi, L., Forbi, J.C., Yokosawa, J., Zelikovsky, A., Khudyakov, Y., 2012. Efficient error correction for next-generation sequencing of viral amplicons. BMC Bioinform. 13 (10), 1–13, http://dx.doi.org/10.1186/1471-2105-13-S10-S6.
Skums, P., Mancuso, N., Artyomenko, A., Tork, B., Mandoiu, I., Khudyakov, Y., Zelikovsky, A., 2013. Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows. BMC Bioinform. 14 (9), 1–13, http://dx.doi.org/10.1186/1471-2105-14-S9-S2.
Töpfer, A., Zagordi, O., Prabhakaran, S., Roth, V., Halperin, E., Beerenwinkel, N., 2013. Probabilistic inference of viral quasispecies subject to recombination. J. Comput. Biol. 20 (2), 113–123, http://dx.doi.org/10.1089/cmb.2012.0232.
Töpfer, A., Marschall, T., Bull, R.A., Luciani, F., Schnhuth, A., Beerenwinkel, N., 2014. Viral quasispecies assembly via maximal clique enumeration. PLoS Comput. Biol. 10 (3), 1–10, http://dx.doi.org/10.1371/journal.pcbi.1003515.
Tsibris, A.M.N., Korber, B., Arnaout, R., Russ, C., Lo, C.-C., Leitner, T., Gaschen, B., Theiler, J., Paredes, R., Su, Z., Hughes, M.D., Gulick, R.M., Greaves, W., Coakley, E., Flexner, C., Nusbaum, C., Kuritzkes, D.R., 2009. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy *in vivo*. PLoS ONE 4 (5), 1–12, http://dx.doi.org/10.1371/journal.pone.0005683.
Vandenhende, M.A., Bellecave, P., Recordon-Pinson, P., Reigadas, S., Bidet, Y., Bruyand, M., Bonnet, F., Lazaro, E., Neau, D., Fleury, H., Dabis, F., Morlat, P., Masquelier, B., 2014. Prevalence and evolution of low frequency HIV drug resistance mutations detected by ultra deep sequencing in patients experiencing first line antiretroviral therapy failure. PLoS ONE 9 (1), e86771, http://dx.doi.org/10.1371/journal.pone.0086771.
Verbist, B.M., Thys, K., Reumers, J., Wetzels, Y., Van der Borght, K., Talloen, W., Aerssens, J., Clement, L., Thas, O., 2015. VirVarSeq: a low-frequency virus variant detection pipeline for illumina sequencing using adaptative base-calling accuracy filtering. Bioinformatics 31 (1), 94–101, http://dx.doi.org/10.1093/bioinformatics/btu587.
Verbist, B., Clement, L., Reumers, J., Thys, K., Vapirev, A., Talloen, W., Wetzels, Y., Meys, J., Aerssens, J., Bijnens, L., Thas, O., 2015. ViVaMBC: estimating viral sequence variation in complex populations from illumina deep-sequencing data using model-based clustering. BMC Bioinform. 16, 59, http://dx.doi.org/10.1186/s12859-015-0458-7.
Vignuzzi, M., Stone, J.K., Arnold, J.J., Cameron, C.E., Andino, R., 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439, 344–348, http://dx.doi.org/10.1038/nature04388.
Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M., Shafer, R.W., 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res. 17 (8), 1195–1201, http://dx.doi.org/10.1101/gr.6468307.
Warren, R.L., Sutton, G.G., Jones, S.J.M., Holt, R.A., 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23 (4), 500–501, http://dx.doi.org/10.1093/bioinformatics/btl629.
Westbrooks, K., Astrovskaya, I., Campo, D., Khudyakov, Y., Berman, P., Zelikovsky, A., 2008. HCV quasispecies assembly using network flows. In: Bioinformatics Research and Applications: Fourth International Symposium, ISBRA 2008. Proceedings, Atlanta, GA, USA, May 6–9, 2008. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 159–170, http://dx.doi.org/10.1007/978-3-540-79450-9_15.
Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., Nagarajan, N., 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucl. Acids Res. 40 (22), 11189–11201, http://dx.doi.org/10.1093/nar/gks918.
Wirden, M., Malet, I., Derache, A., Marcelin, A.G., Roquebert, B., Simon, A., Kirstetter, M., Joubert, L.M., Katlama, C., Calvez, V., 2005. Clonal analyses of HIV quasispecies in patients harbouring plasma genotype with K65R mutation associated with thymidine analogue mutations or L74V substitution. AIDS 19 (6), 630–632.
Woo, H.-J., Reifman, J., 2012. A quantitative quasispecies theory-based model of virus escape mutation under immune selection. Proc. Natl. Acad. Sci. U. S. A. 109 (32), 12980–12985, http://dx.doi.org/10.1073/pnas.1117201109.
Yang, X., Charlebois, P., Gnerre, S., Coole, M.G., Lennon, N.J., Levin, J.Z., Qu, J., Ryan, E.M., Zody, M.C., Henn, M.R., 2012. *De novo* assembly of highly diverse viral populations. BMC Genomics 13 (1), 1–13, http://dx.doi.org/10.1186/1471-2164-13-475.
Yang, X., Charlebois, P., Macalalad, A., Henn, M.R., Zody, M.C., 2013. V-Phaser 2: variant inference for viral populations. BMC Genomics 14, 674, http://dx.doi.org/10.1186/1471-2164-14-674.
Yoon, B.-J., 2009. Hidden Markov models and their applications in biological sequence analysis. Curr. Genomics 10 (6), 402–415, http://dx.doi.org/10.2174/138920209789177575.
Zagordi, O., Geyrhofer, L., Roth, V., Beerenwinkel, N., 2010. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. J. Comput. Biol. 17 (3), 417–428, http://dx.doi.org/10.1089/cmb.2009.0164.

Zagordi, O., Klein, R., Däumer, M., Beerenwinkel, N., 2010. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucl. Acids Res. 38 (21), 7400–7409, http://dx.doi.org/10.1093/nar/gkq655.

Zagordi, O., Bhattacharya, A., Eriksson, N., Beerenwinkel, N., 2011. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinform. 12, 119, http://dx.doi.org/10.1186/1471-2105-12-119.

Zagordi, O., Däumer, M., Beisel, C., Beerenwinkel, N., 2012. Read length versus depth of coverage for viral quasispecies reconstruction. PLoS ONE 7 (10), 1–8, http://dx.doi.org/10.1371/journal.pone.0047046.

Zanini, F., Brodin, J., Thebo, L., Lanz, C., Bratt, G., Albert, J., Neher, R.A., 2015. Population genomics of intrapatient HIV-1 evolution. eLife 4, e11282, http://dx.doi.org/10.7554/eLife.11282.

Zeng F., Jiang R., Ji G. Chen T. ProbAlign: a re-alignment method for long sequencing reads. bioRxivdoi:10.1101/008698.